

# Nonparametric Methods for Sample Surveys of Environmental Populations

*Metodi nonparametrici nell'inferenza per popolazioni finite di carattere ambientale*

Giorgio E. Montanari  
Dipartimento di Economia,  
Finanza e Statistica  
Università degli Studi di Perugia  
giorgio@stat.unipg.it

M. Giovanna Ranalli  
Dipartimento di Economia,  
Finanza e Statistica  
Università degli Studi di Perugia  
giovanna@stat.unipg.it

**Riassunto:** Il lavoro presenta una rassegna della letteratura sull'impiego di tecniche di regressione nonparametrica nell'inferenza assistita dal modello su popolazioni finite di carattere ambientale. Affronta poi il problema della stima della funzione di distribuzione di una variabile oggetto di indagine in presenza di informazione ausiliaria completa. I pesi di riporto all'universo sono determinati attraverso il metodo di massima pseudo-verosimiglianza empirica vincolata. I vincoli garantiscono, da una parte, pesi che forniscono una funzione di distribuzione propria, dall'altra, l'impiego dell'informazione ausiliaria attraverso un modello nonparametrico molto generale. Il metodo proposto è infine applicato alla stima dell'acidità dei laghi nel Nord-Est degli Stati Uniti.

**Keywords:** Auxiliary information; Nonparametric regression; Pseudo empirical likelihood; Model-assisted approach; MARS.

## 1. Introduction

Auxiliary information to be used to increase the accuracy of estimates of finite population parameters has become fairly common. This is certainly true for social, economic and demographic surveys where information coming from previous surveys, administrative registers and census data can be employed at the estimation stage. Such information can be incorporated in a *model-assisted* approach to inference (Särndal *et al.*, 1992). In the latter, a superpopulation model describing the relationship between the variable of interest and the auxiliary variables is used to construct sample-based estimators that are efficient when the model is correct, but maintain key design properties such as design consistency when the model is incorrect. The increase in efficiency compared to the unbiased Horvitz-Thompson estimator relies on the accuracy and complexity of the superpopulation model employed. This, in turn, is also related to the type of auxiliary information available. In particular, when the values of a set of auxiliary variables are known for all units in the population, we have the so called *complete* auxiliary information. When this is not true, auxiliary information is in the form of population level means, totals or counts of the auxiliary variables. In the latter case essentially only linear regression models can be employed, while in the former virtually any type of statistical model can be applied: linear, nonlinear, generalized linear, nonparametric regression models.

Complex sampling designs have been recently employed also for environmental studies. The need of surveying environmental resources as surface waters and forests has developed a survey sampling framework to address the issue of assessing their ecological

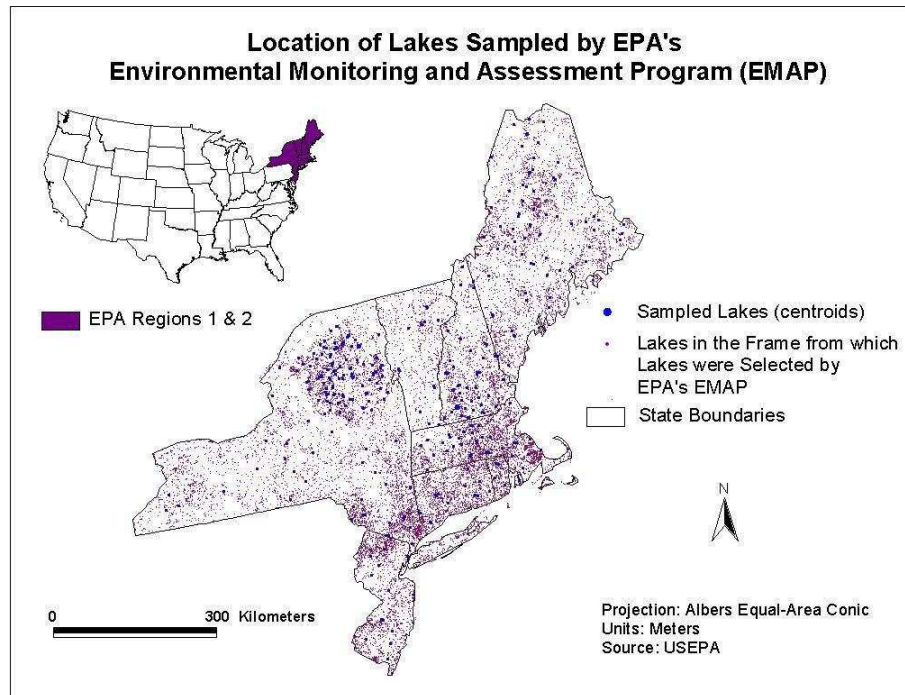


Figure 1: Map of frame and survey lakes locations from the EMAP study on the ecological condition of lakes in northeastern U.S.

condition. As an illustration, consider the lakes in the northeastern region of the United States reported in Figure 1. The National Surface Water Survey sponsored by the United States Environmental Protection Agency (EPA) between the years of 1984 and 1986 estimated 4.2 percent of the lakes to be acidic (Stoddard *et al.*, 2003). These acid-sensitive Northeastern lakes were among the concerns addressed by the Clean Air Act Amendment (CAAA) issued by EPA in 1990, which placed restrictions on industrial sulfur and nitrogen emissions in an effort to reduce the acidity of these waters. Between 1991 and 1996, the Environmental Monitoring and Assessment Program (EMAP) of EPA conducted a survey of lakes in the Northeastern states of the U.S. These data were collected in order to determine the effect that restrictions put in place by the CAAA had on the ecological condition of these waters. The survey is based on a population of 21,026 lakes from which 334 lakes were surveyed, some of which were visited several times during the study period (see Figure 1 for a map of both frame and sample locations). Lakes to be included in the survey were selected using a complex sampling design commonly employed by EMAP. It is based on a hexagonal grid frame and assigns inclusion probabilities to lakes according to the size class they belong to (Larsen *et al.*, 1993). Estimation of quantities of interest at the population level can be carried on using a model-assisted approach. Models can be built relying on complete auxiliary information in the form of spatially-referenced data maintained in a geographic information system (GIS). Satellite images, in fact, can provide the values of variables thought to influence the process under study for each frame location: land cover, ecosystem typology, elevation can be obtained at little or no extra cost from GIS maps.

Nonparametric regression methods have been employed in an environmental survey sampling context ever after the theoretical pioneering work of Breidt and Opsomer (2000)

on local polynomials regression estimators. Kim *et al.* (2004) consider such estimator to estimate finite population totals in two-stage sampling for which complete auxiliary information is available for first-stage sampling units. This method is applied to data from the 1995 National Resources Inventory Erosion Update Study. The National Resources Inventory (NRI) is a stratified two-stage area sample of the agricultural lands in the United States conducted by the Natural Resources Conservation Service of the U.S. Department of Agriculture. The 1995 Erosion Update Study was a smaller-scale study that used NRI information as frame material. The auxiliary variable employed to estimate totals of wind and water erosion is the size measure of land with erosion potential.

Local polynomials are mostly suitable to handle univariate auxiliary information; applicability of nonparametric regression methods has been extended to multivariate auxiliary information by use of more appropriate nonparametric techniques. Opsomer *et al.* (2001) and Opsomer *et al.* (2005) deal with a multi-phase survey conducted for a forest inventory in the mountains of Utah in the United States. For such survey in phase one, remote sensing data and GIS coverage information are extracted on an intensive sample grid. Phase two consists of a field-visited subset of the phase one grid. Generalized additive models are used in a model-assisted approach to estimate totals of variables regarding tree characteristics, size measurements, ratings of ecological health collected on the field visits and accounting for the two levels of auxiliary information. Breidt *et al.* (2005) also address this issue and apply penalized splines (Ruppert *et al.*, 2003). This nonparametric technique is gaining much popularity for its flexibility and ease of implementation; in this framework, further, it allows a complete treatment of theoretical design properties which cannot be developed in the case of Generalized additive models. Penalized splines are also natural candidates to introduce a nonparametric trend in a small area estimation context (Opsomer *et al.*, 2004); it is, in fact, possible to express the estimation problem as a mixed effect model regression and obtain a mean estimate of Acid Neutralizing Capacity (ANC) in the northeastern lakes of Figure 1 for each of 113 small areas defined by 8-digit Hydrologic Unit Codes within the region of interest. Salvati (2005) employs M-quantile regression for small area estimation. A nonparametric generalization of model calibration introduced by Wu and Sitter (2001) has been implemented with neural networks and local polynomials in Montanari and Ranalli (2005) to estimate the population mean of Total Nitrogen and Total Phosphorus concentrations in the streams surveyed in the Mid-Atlantic Highlands of the United States. The proportion of land devoted to agriculture in a particular watershed has been used as auxiliary covariate.

In this work we explore the possibility of employing nonparametric techniques, and in particular Multi Adaptive Regression Splines (MARS, Friedman, 1991) to build a model-assisted estimator of the distribution function (cdf) of ANC in the northeastern lakes survey. ANC is a common measure of acidity defined as a water's ability to buffer acid. Here, in fact, concern is mainly with the assessment of how many lakes are at (high) risk of acidification or are acidified already. The estimator of the cdf and the corresponding confidence intervals are based on Model Calibrated Pseudo Empirical Maximum Likelihood (MCPEML) estimators proposed in Chen and Wu (2002) and Wu and Rao (2005). Nonparametric model calibration has been introduced in Montanari and Ranalli (2005) and used to estimate totals and means also for environmental populations; although it could be applied as is to cdf estimation, it would have the drawback of possibly taking values outside the interval  $[0, 1]$  and of not always being a monotone function of the response variable. Therefore, if on one side nonparametric regression allows a more flexible modeling of ANC with respect to remote sensed auxiliary variables, on the other side

MCPEML assures the achievement of a genuine distribution function. Finally, the estimation of confidence intervals through pseudo empirical likelihood has been shown to be superior over normal confidence bounds (Wu and Rao, 2005) especially for cdf estimation.

The work proceeds as follows. Section 2.1 introduces some notation and revises model calibration. Pseudo empirical maximum likelihood (PEML) estimation is considered in Section 2.2, while nonparametric regression for cdf estimation in this context is introduced in Section 2.3. The issue of estimating confidence intervals is addressed in Section 2.4. Section 3 shows the results of the application of these techniques to cdf estimation of ANC for the northeastern lakes survey. Some concluding remarks are given in Section 4.

## 2. Nonparametric MCPEML estimation

### 2.1 Model calibration

Consider a finite population  $\mathcal{U} = \{1, \dots, N\}$ . For each unit in the population we assume that the value of a vector  $\mathbf{x}$  of  $Q$  auxiliary variables is available and therefore the vector  $\mathbf{x}_i = (x_{1i}, \dots, x_{qi}, \dots, x_{Qi})$  is known  $\forall i \in \mathcal{U}$ . A sample  $s$  of size  $n$  is drawn from  $\mathcal{U}$  according to a probabilistic sampling plan with inclusion probabilities  $\pi_i$  and  $\pi_{ij}$ , for all  $i, j \in \mathcal{U}$ . The survey variable  $y$  is observed for each unit in the sample and the goal here is to estimate the population distribution function of the survey variable, that is  $F_N(t) = N^{-1} \sum_{i \in \mathcal{U}} I(y_i \leq t)$ . The population cdf can be itself seen as a population mean of the indicator variable  $z_i = I(y_i \leq t)$ , so that without using any auxiliary information the following ratio estimator is the well established Hajek estimator:

$$\hat{F}_H(t) = \frac{\sum_{i \in s} d_i I(y_i \leq t)}{\sum_{i \in s} d_i} = \sum_{i \in s} d_i^* I(y_i \leq t), \quad (1)$$

with  $d_i = 1/\pi_i$  and  $d_i^* = d_i / \sum_{i \in s} d_i$ . In the presence of auxiliary information, straightforward application of techniques developed for the estimation of the population mean  $\bar{y}_N = N^{-1} \sum_{i \in \mathcal{U}} y_i$  of  $y$  can be misleading. A simple example is a regression-type estimator for  $F_N(t)$  that would have the form  $\hat{F}_R(t) = \hat{F}_H(t) + (F_{N\hat{y}}(t) - \hat{F}_{H\hat{y}}(t))$ , where  $F_{N\hat{y}}(t)$  is the population cdf of the fitted values  $\hat{y}_i = \mathbf{x}_i \mathbf{B}_x$ ,  $\hat{F}_{H\hat{y}}(t)$  is its Hajek-type estimator and  $\mathbf{B}_x$  is the vector of estimated regression coefficients in a generalized-type regression of  $y_i$  on  $\mathbf{x}_i$ . This type of estimator is not a distribution function and can therefore take values outside the interval  $[0, 1]$ .

A generalized regression estimator of  $\bar{y}_N$  can be seen as an important particular case of calibration estimators as dealt with in Deville and Särndal (1992): it is a weighted mean of sample values of  $y$ , with weights  $w_i$  that minimize the distance  $\sum_{i \in s} (w_i - d_i)^2 / d_i$  from the basic design weights, while meeting benchmark constraints that ensure internal consistency with the auxiliary information on the  $\mathbf{x}$  variables. In fact, weights  $w_i$  provide perfect estimates when applied to the auxiliary variables, in the sense that the mean estimate of  $\mathbf{x}$  takes the known value of the population mean. This requirement is not generally needed when estimating  $F_N(t)$ . This would set as a natural candidate for the estimation of  $F_N(t)$  the generalization of calibration estimation proposed in Wu and Sitter (2001) and also studied in Montanari and Ranalli (2005): *model calibration* first adopts a superpopulation model – either parametric or nonparametric – to describe the relationship between survey and auxiliary variables and then calibrates over the fitted values obtained from the model.

Suppose the relationship between  $y$  and  $\mathbf{x}$  can be well described through the following regression model:

$$E_\xi(y_i|\mathbf{x}_i) = \mu(\mathbf{x}_i), \quad V_\xi(y_i|\mathbf{x}_i) = v(\mathbf{x}_i), \quad \text{for } i = 1, \dots, N, \quad (2)$$

where  $E_\xi$  and  $V_\xi$  denote the expectation and the variance with respect to the model,  $\mu(\cdot)$  and  $v(\cdot)$  are known functions of unknown parameters (Wu and Sitter, 2001) or unknown smooth functions (Montanari and Ranalli, 2005). Calibration is then performed on the population mean of design based estimates of  $\mu(\mathbf{x}_i)$ . This allows a more efficient use of complete auxiliary information, but, when applied to the issue of cdf estimation, would still suffer from the drawback of providing estimates that might take values outside the allowed range.

## 2.2 Pseudo empirical maximum likelihood estimation

The PEML estimator of  $\bar{y}_N$  will be defined as  $\hat{y}_{\text{PEML}} = \sum_{i \in s} \hat{p}_i y_i$ , with weights  $\hat{p}_i$  obtained as the maximizers of the following pseudo empirical log-likelihood function

$$l_n(\mathbf{p}) = n^* \sum_{i \in s} d_i^* \log(p_i) \quad (3)$$

subject to the set of constraints

$$0 < p_i < 1, \quad \sum_{i \in s} p_i = 1, \quad \sum_{i \in s} p_i \mathbf{g}_i = \bar{\mathbf{g}}_N, \quad (4)$$

where  $\mathbf{g}_i$  is a set of known functions of the auxiliary variables,  $\bar{\mathbf{g}}_N = N^{-1} \sum_{i \in \mathcal{U}} \mathbf{g}_i$  is its known population mean and  $n^*$  is the effective sample size, a quantity related to the design effect whose details will be given in what follows. The original pseudo empirical likelihood function proposed in Chen and Sitter (1999) is  $l(\mathbf{p}) = \sum_{i \in s} d_i \log(p_i)$ ; the resulting estimator is the same using either of the two, but the rescaling employed in (3) will be useful when addressing construction of the confidence intervals for the estimator. If in (4)  $\mathbf{g}_i = \mathbf{x}_i$ , the resulting estimator is asymptotically equivalent to a generalized regression estimator (Chen and Sitter, 1999). Wu and Sitter (2001) extend this approach to model calibration and employ the scalar  $g_i = \mu(\mathbf{x}_i)$  in (4). Wu (2003) shows that the resulting MCPML estimator is optimal among the class of PEML estimators, in that the expected value of the asymptotic design variance under the model and any regular sampling design with fixed sample size reaches its minimum. The Lagrange multiplier method can be used to show that in this case

$$\hat{p}_i = \frac{d_i^*}{1 + \lambda(g_i - \bar{g}_N)}, \quad \text{for } i \in s \quad (5)$$

and the scalar Lagrange multiplier  $\lambda$  is the solution to

$$\sum_{i \in s} \frac{d_i^*(g_i - \bar{g}_N)}{1 + \lambda(g_i - \bar{g}_N)} = 0. \quad (6)$$

Clearly, in applications  $\mu(\mathbf{x}_i)$  will be replaced by its design based estimates; for parametric models, estimates of the model parameters can be obtained by means of weighted

estimating equations (Wu and Sitter, 2001), while in the nonparametric case, design based adjustments must be performed according to the method employed (e.g. Breidt and Opsomer, 2000; Breidt *et al.*, 2005; Montanari and Ranalli, 2005; Opsomer *et al.*, 2005). A modified Newton-Raphson algorithm to find a solution to (6) has been proposed in Chen *et al.* (2002) and R functions that implement it can be found in Wu (2005). The resulting MCPPEML estimator for  $\bar{y}_N$  is asymptotically equivalent to the ordinary model calibration estimator and therefore shares its design-based properties as consistency. However, for finite samples a very attractive feature of this estimator relies in the intrinsic properties of the weights  $\hat{p}_i$  – that is  $\hat{p}_i > 0$  and  $\sum_{i \in s} \hat{p}_i = 1$  – which are particularly valuable when estimating the distribution function.

### 2.3 MCPPEML for cdf estimation

To estimate  $F_N(t)$  for a given  $t_0$ , we need to replace  $y_i$  in the preceding developments by  $z_i = I(y_i \leq t_0)$  and choose values  $g_i$  to put in (4). The choice  $g_i = E_\xi(z_i | \mathbf{x}_i) = P(y_i \leq t_0 | \mathbf{x}_i)$  is optimal in the sense described earlier (Wu, 2003). Note that this choice depends on  $t_0$ ; therefore no  $g_i$  with a fixed  $t_0$  can be uniformly optimal for  $F_N(t)$  for all values of  $t$  (Chen and Wu, 2002). We will also see in the application the consequences of using a single set of weights for all  $t$ : although this might not be the most efficient solution, this procedure results in a genuine distribution function that can be used also for quantile estimation, and is shown to be very efficient for values of  $t$  in a wide neighborhood of  $t_0$ .

Since the response variable is now the indicator variable  $z_i$ , there are two types of working models that can be considered to obtain  $g_i$  values: models that relate the  $y_i$  to the  $\mathbf{x}_i$  or models that relate the indicators  $z_i$  to the  $\mathbf{x}_i$ . In the first case model (2) would be assumed and the  $g_i$  values to be used in (4) would be given by

$$g_i = P(y_i \leq t | \mathbf{x}_i) = G\{(t - \mu(\mathbf{x}_i))/v(\mathbf{x}_i)\},$$

where  $G(\cdot)$  is the cdf of the error component  $\varepsilon_i = (y_i - \mu(\mathbf{x}_i))/v(\mathbf{x}_i)$ . When the normality assumption for the distribution of the errors can be made, then the cdf of a standard normal distribution can be employed for  $G(\cdot)$ . When such an assumption is not desirable for  $G(\cdot)$ , then this function has to be estimated from fitted residuals (Chen and Wu, 2002). Here the indicators  $z_i$  are modeled indirectly through a model for the  $y_i$ . A more attractive solution, though, is given by exploiting the probability nature of the  $g_i$ . A generalized-type model can be therefore employed; Chen and Wu (2002) and Wu (2003) explore this possibility in the form of a logistic regression model for the  $z_i$ ; i.e.  $\log(g_i/(1 - g_i)) = \mathbf{x}_i\beta$ , with variance function given by  $V_\xi(g) = g(1 - g)$ . One of the advantage of using such a model is that the error distribution in the regression model is no longer an issue. In this paper, we would like to extend this definition to more general models that can better accomodate auxiliary information. We will consider the following model for the indicators  $z_i$ :

$$\log\left(\frac{g_i}{1 - g_i}\right) = \mu(\mathbf{x}_i), \quad (7)$$

where  $\mu(\mathbf{x}_i)$  is again the very general function considered in (2). If such function is left undefined, nonparametric techniques have to be employed to estimate it. In the application we will explore the applicability of MARS (Friedman, 1991). In general, once design based estimates  $\hat{\mu}_i = \hat{\mu}(\mathbf{x}_i)$  of  $\mu(\mathbf{x}_i)$  are obtained – via generalized estimating equations or weight adjusted nonparametric techniques – the scalars  $g_i$  to be employed in the set of constraints (4) will be given by  $g_i = \exp(\hat{\mu}_i)/\{1 + \exp(\hat{\mu}_i)\}$ . The set of MCPPEML weights will be then obtained by using the Newton-Rapshon algorithm in Chen *et al.* (2002).

## 2.4 Confidence intervals

Wu and Rao (2005) establish the asymptotic distribution of the pseudo empirical likelihood ratio statistics related to (3) under certain regularity conditions and in an asymptotic framework as that of Isaki and Fuller (1982), and obtain the associated confidence intervals for  $F_N(t)$  at  $t = \tilde{t}$ , say, based on these results. Their findings use auxiliary information in the form of known population means of  $\mathbf{x}$  and therefore employ  $\mathbf{g}_i = \mathbf{x}_i$  in the set of constraints (4). Generalization to more complex models is possible as long as the fitted values  $\hat{\mu}_i$  fulfill the regularity requirements on their asymptotic behavior required there for  $\mathbf{x}_i$ .

Since we will compute the Hajek estimator for comparison matters in Section 3, we will first describe the PEML confidence intervals in the case of no auxiliary information at the estimation stage and then move on to the more complex case. With no auxiliary information, maximizing  $l_n(\mathbf{p})$  in (3) subject to  $p_i > 0$  and  $\sum_{i \in s} p_i = 1$  gives  $\hat{p}_i = d_i^*$  and, therefore, the Hajek estimator. Let  $\tilde{p}_i$  be the value of  $p_i$  obtained by maximizing  $l_n(\mathbf{p})$  subject to  $p_i > 0$ ,  $\sum_{i \in s} p_i = 1$  and  $\sum_{i \in s} p_i z_i = \theta$ , for  $z_i = I(y_i \leq \tilde{t})$  and a fixed  $\theta$ . The effective sample size  $n^*$  is defined to be

$$n^* = \frac{S_z^2}{V_p(\sum_{i \in s} d_i^* z_i)}, \quad (8)$$

where  $S_z^2 = (N - 1)^{-1} \sum_{i \in \mathcal{U}} (z_i - F_N(\tilde{t}))^2$  is the population variance of the indicator variables and  $V_p$  is the design-based variance of the Hajek estimator. Wu and Rao (2005) prove that the ratio statistics  $r_n(\theta) = -2\{l_n(\tilde{\mathbf{p}}) - l_n(\hat{\mathbf{p}})\}$  converges in distribution to a  $\chi_1^2$  random variable when  $\theta = F_N(\tilde{t})$ . As a consequence, the  $1 - \alpha$  PEML confidence interval for  $F_N(\tilde{t})$  will be given by the set  $\{\theta | r_n(\theta) < \chi_1^2(\alpha)\}$ , where  $\chi_1^2(\alpha)$  is the  $1 - \alpha$  quantile of a  $\chi^2$  distribution with one degree of freedom. The effective sample size  $n^*$  will have to be estimated for practical applications for each  $t$ .

Let us now consider the case of presence of complete auxiliary information at the estimation stage. Such information is employed through a working model of the type in (7). Therefore, the extra benchmark constraint  $\sum_{i \in s} p_i g_i = \bar{g}_N$  is employed to obtain both  $\hat{\mathbf{p}}$  and  $\tilde{\mathbf{p}}$ . Let again  $z_i = I(y_i \leq \tilde{t})$  and let  $n^*$  in (8) be replaced by

$$n^* = \frac{S_e^2}{V_p(\sum_{i \in s} d_i^* e_i)}, \quad (9)$$

where  $e_i = z_i - Bg_i$  is a residual variable of the population level regression of  $z_i$  on  $g_i$  and  $B$  is the coefficient of such regression. The asymptotic approximation of the ratio statistic to a  $\chi_1^2$  distribution still holds; estimates of  $n^*$  must be computed for practical applications for each  $t$ , while  $g_i$  are estimated only for a fixed point  $t = t_0$  to obtain a genuine cdf (Section 2.3).

## 3. Assessment of the ecological condition of the lakes in Northeastern U.S. through the estimation of the ANC cdf

In this section we apply the aforementioned technique to estimate the distribution function of ANC in the northeastern lakes of the U.S. Recall that the survey is based on a population of 21,026 lakes from which 334 lakes were surveyed (Figure 1), some of which were visited several times during the study period. The total number of measurements is

$t$	Hajek		MARS	
	$\hat{F}_H(t)$	95% CI	$\hat{F}_{\text{MARS}}(t)$	95% CI
0	0.060	(0.017; 0.143)	0.060	(0.017; 0.140)
50	0.164	(0.104; 0.238)	0.162	(0.103; 0.234)
200	0.411	(0.301; 0.527)	0.408	(0.311; 0.505)

Table 1: Cdf estimates at  $t = 0, 50, 200$  and relative 95% confidence intervals by the Hajek and Mars estimators. The average length of the confidence intervals is 0.162 with  $\hat{F}_H(t)$  and 0.149 with  $\hat{F}_{\text{MARS}}(t)$ ;  $\hat{F}_{\text{MARS}}(t)$  was computed fixing  $t_0 = 200$ .

551; if multiple measurements are available for the same lake, we average these in order to obtain one measurement per lake sampled. Since the inclusion probability of each lake is determined according to its size class, a  $\pi$ ps-sampling is a good approximation to the complex sampling design based on a hexagonal grid frame employed by EMAP.

Let  $y_i$  represent the (possibly averaged) ANC value of the  $i$ -th sampled lake,  $i = 1, \dots, 334$ . An ANC value less than 0  $\mu\text{eq/L}$  indicates that the water has lost all ability to buffer acid. Surface waters with ANC values below 200  $\mu\text{eq/L}$  are considered at risk of acidification, and values less than 50  $\mu\text{eq/L}$  are considered at high risk. We will therefore, first obtain estimates and the relative confidence intervals of  $F_N(t)$  at these three values of  $t$  because of their importance and we will then move to the estimation of the whole  $F_N(t)$ .

In the case of no auxiliary information,  $\hat{F}_H(t)$  can be computed together with PEML confidence intervals as described in Sections 2.1 and 2.4. To compute the confidence intervals, we need to estimate  $n^*$  in (8). The numerator requires estimation of the population variance of the indicator variable. Courbois and Urquhart (2004) explore this issue and relate the choice among different estimators (weighted, non-weighted, ratio-type) to the correlation between the response variable and the inclusion probabilities, and to the inclusion probabilities variance. In our case, the correlation between the indicator variables at  $t = 0, 50, 200$  and the inclusion probabilities takes really small values ranging from almost zero to 0.25. On the other side, the variance of the inclusion probabilities relative to its possible maximum under  $\pi$ ps sampling (Courbois and Urquhart, 2004, equation 3.1) takes value 0.08. Courbois and Urquhart (2004) suggest the use of the naive sample variance over weighted and ratio-type estimators when both the correlation between the response variable and the inclusion probabilities, and the variance of the inclusion probabilities take small values. We will follow this advice.

The denominator of (8) requires the estimation of the design variance of the Hajek estimator. This involves second order inclusion probabilities (see e.g. Särndal *et al.*, 1992, p.182), which are not available from this survey. We therefore use an approximation suggested by Stehman and Overton (1989) for which  $\hat{\pi}_{ij} = \frac{(n-1)\pi_i\pi_j}{2n-\pi_i-\pi_j}$ , with  $\pi_i$  substituted for  $\hat{\pi}_{ii}$ . The final estimates of  $\hat{F}_H(t)$  for  $t = 0, 50, 200$  and relative 95% confidence intervals are reported in Table 1. Computation has been carried with the aid of the R functions provided in Wu (2005).

Auxiliary variables are available for each lake in this population; this should make it possible to improve upon the efficiency of the Hajek estimator. The following variables are available for each  $i \in U$ :



- $x_{1i}$  = UTMX,  $x$ -geographical coordinate of the centroid of each lake in the UTMcoordinate system,
- $x_{2i}$  = UTMY,  $y$ -geographical coordinate,
- $x_{3i}$  = categorical variable for eco-region (7 levels),
- $x_{4i}$  = elevation.

We employ model (7) and approximate the unknown smooth function  $\mu(\mathbf{x}_i)$  through MARS. Employing a nonparametric model for these data was also suggested by evidence in Opsomer *et al.* (2004) of a bivariate surface in the geographical coordinates when estimating ANC means for small areas. MARS also allows to detect both interactions and nonlinearities without making unduly restrictive a priori assumptions and provides interpretable results unlike other yet powerful techniques as neural networks. The collection of Fortran subroutines MARS 3.6 has been employed to fit MARS to the sample data; no use of the basic design weights has been made within the fitting procedure. Recall the discussion at the beginning of Section 2.3: although no  $g_i$  with a fixed  $t_0$  can be uniformly optimal for  $F_N(t)$  at all values of  $t$ , this is a requirement to obtain a genuine distribution function. We therefore employ the value  $t_0 = 200$  and obtain fitted values only for the model that relates  $I(y_i \leq 200)$  to  $\mathbf{x}_i$ . The value 200 has been chosen so that it could be used also to get the whole cdf as shown hereafter. There is no general guideline to choose the value  $t_0$  apart from covering more efficiently a neighborhood of interest. Model selection through the generalized cross validation criterion used in Friedman (1991) determines that 15 candidate basis functions are to be used. The estimates  $\hat{\mu}_i$  are employed to obtain the  $g_i$  to be used in the constraints (4). This leads us to the set of weights we use for the three values of  $t$ . These  $g_i$  are also employed to calculate the generalized regression-type estimate of the regression coefficient  $B$  and the residuals employed to get  $n^*$  in (9). The final estimates of  $\hat{F}_{\text{MARS}}(t)$  for  $t = 0, 50, 200$  and relative 95% confidence intervals are reported in Table 1.

The estimates of the cdf are comparable in all cases. The confidence intervals with  $\hat{F}_H(t)$  are on average about 9% wider than those with  $\hat{F}_{\text{MARS}}(t)$ . Based on these results we could try to evaluate the effect that emissions restrictions put in place by the CAAA of 1990 have had on levels of acidity in these lakes. Recall that a previous survey determined 4.2% of Northeastern Lakes to be acidic in 1986. Based on EMAP survey data taken between the years of 1991 and 1996, both  $\hat{F}_{\text{MARS}}(t)$  and  $\hat{F}_H(t)$  estimate 6% of Northeastern lakes to be acidic. These estimates do not show any evidence of a reduction of acid levels of waters in the Northeastern region of the United States. However, more recent data are needed to assess whether the ANC of Northeastern lakes may have a delayed response to the emission restrictions to show significant reductions.

Figure 2 shows the cdf estimate performed by  $\hat{F}_H(t)$  and  $\hat{F}_{\text{MARS}}(t)$  with the associate confidence intervals. The estimates with MARS were again computed fixing  $t_0 = 200$ . Here the difference in the width of the confidence intervals is more striking in favor of  $\hat{F}_{\text{MARS}}(t)$ ; confidence intervals associated with  $\hat{F}_H(t)$  are, on average, almost 40% wider than those associated with  $\hat{F}_{\text{MARS}}(t)$ . Figure 3 shows the curves and surfaces estimated by MARS in the model for  $I(y_i \leq 200)$  with 15 candidate basis functions. All variables turned out to be important, in that their elimination would have increased the generalized cross validation index considerably. Model selection provided evidence of the inclusion of the variable  $x_4$  (elevation) as an additive variable, i.e. interactions with other variables were not improving in the model. This is a sensitive results, as well as the shape of the final curve estimated for this variable (see the first panel of Figure 3). In fact, it makes

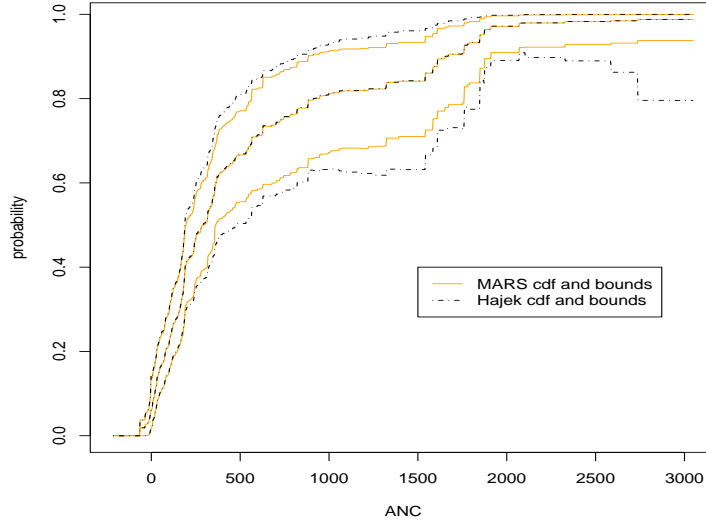


Figure 2: Estimates of the ANC cdf and relative confidence bounds produced by  $\hat{F}_H(t)$  and  $\hat{F}_{\text{MARS}}(t)$  for a 1000 grid values. The average length of the confidence intervals is 0.209 with  $\hat{F}_H(t)$  and 0.149 with  $\hat{F}_{\text{MARS}}(t)$ ;  $\hat{F}_{\text{MARS}}(t)$  was computed fixing  $t_0 = 200$ .

sense that small values of ANC are associated with large values of elevation, which in turn are associated with fresher waters. In fact, acid falls in the form of rain and, as the water flows, the calcium in the limestone bedrock buffers the acid and increases ANC. The second panel shows an interaction between the  $x$  and  $y$ -coordinate of the lake location; it is important to note that this surface does not represent a smooth of ANC on the  $x$  and  $y$ -coordinate, but rather it shows the contribution of the  $x$  and  $y$ -coordinate to the smooth predictor  $\hat{\mu}_i$  on the four variables employed.

#### 4. Concluding remarks

A review of applications of nonparametric regression to model-assisted inference for environmental populations is provided. When auxiliary information is available for all units in the population, nonparametric techniques have been shown to be useful tools to improve in terms of efficiency over Horvitz-Thompson and regression-type estimators. In this work, the issue of estimating the distribution function of ANC – a measure of acidity of surface waters – from a survey of lakes in Northeastern U.S. is addressed by making use of remotely sensed auxiliary information available at the frame level. Nonparametric model calibration applied to pseudo empirical maximum likelihood is employed to obtain a set of modified design weights that use the geographical coordinates of the lake's location, its elevation and type of eco-region by means of a logistic MARS model. Gains in estimated efficiency are shown over the Hajek estimator in terms of less wide confidence intervals. The extremes of the confidence intervals are estimated exploiting the pseudo empirical maximum likelihood nature of the modified design weights.

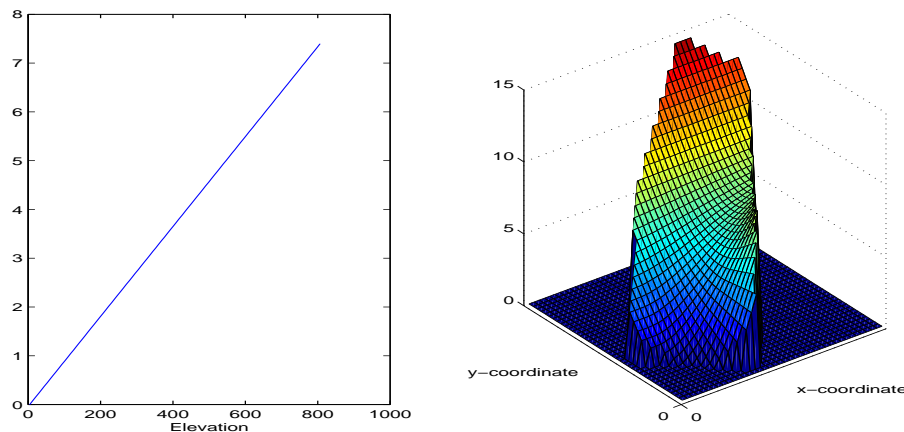


Figure 3: Curves and surfaces estimated by MARS in the model for  $t_0 = 200$ . The vertical axis in each plot shows the contribution of each variable to the whole smooth predictor  $\hat{\mu}_i$ ; since the locations of the plotted functions are arbitrary, they are all translated to have zero minimum value.

## Acknowledgements

The work reported here was conducted while the second author was appointed at Colorado State University and supported by STAR Research Assistance Agreements CR-829095 and CR-829096 awarded by the U.S. Environmental Protection Agency (EPA) to Colorado State University and Oregon State University. This manuscript has not been formally reviewed by EPA. The views expressed here are solely those of the authors. EPA does not endorse any products or commercial services mentioned in this report.

## References

- Breidt F.J., Caleskens G. and Opsomer J.D. (2005) Model-assisted estimation for complex surveys using penalized splines, *Working Paper*, Preprint Series 03–15, Department of Statistics, Iowa State.
- Breidt F.J. and Opsomer J.D. (2000) Local polynomial regression estimators in survey sampling, *The Annals of Statistics*, 28, 1026–1053.
- Chen J. and Sitter R.R. (1999) A pseudo-empirical likelihood approach to the effective use of auxiliary information in complex surveys, *Statistica Sinica*, 9, 385–406.
- Chen J., Sitter R.R. and Wu C. (2002) Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys, *Biometrika*, 89, 1, 230–237.
- Chen J. and Wu C. (2002) Estimation of distribution function and quantiles using the model-calibrated pseudo empirical likelihood method, *Statistica Sinica*, 12, 1223–1239.
- Courbois J.Y. and Urquhart N.S. (2004) Comparison of survey estimates of the finite population variance, *Journal of Agricultural Biological and Environmental Statistics*, 9, 236–251.
- Deville J.C. and Särndal C.E. (1992) Calibration estimators in survey sampling, *Journal of the American Statistical Association*, 87, 376–382.

- Friedman J.H. (1991) Multivariate adaptive regression splines, with discussion, *The Annals of Statistics*, 19, 1–67.
- Isaki C.T. and Fuller W.A. (1982) Survey design under the regression superpopulation model, *Journal of the American Statistical Association*, 77, 89–96.
- Kim J., Breidt F.J. and Opsomer J.D. (2004) Nonparametric regression estimation of finite population totals under two-stage sampling, *Working Paper*, Preprint Series 03–06, Department of Statistics, Iowa State.
- Larsen D.P., Thornton K.W., Urquhart N.S. and Paulsen S.G. (1993) Overview of survey design and lake selection, EMAP – Surface waters 1991 Pilot Report. Technical Report EPA/620/R-93/003, U.S. Environmental Protection Agency.
- Montanari G.E. and Ranalli M.G. (2005) Nonparametric model calibration estimation in survey sampling, *Journal of the American Statistical Association*, in print.
- Opsomer J.D., Breidt F.J., Claeskens G., Kauermann G. and Ranalli M.G. (2004) Nonparametric small area estimation using penalized spline regression, in: *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Alexandria, VA.
- Opsomer J.D., Breidt F.J., Moisen G.G. and Kauermann G. (2005) Model-assisted estimation of forest resources with generalized additive models, *Journal of the American Statistical Association*, in print.
- Opsomer J.D., Moisen G.G. and Kim J.Y. (2001) Model-assisted estimation of forest resources with generalized additive models, in: *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Alexandria, VA.
- Ruppert D., Wand M.P. and Carroll R. (2003) *Semiparametric Regression*, Cambridge University Press, Cambridge, New York.
- Salvati N. (2005) M-quantile geographically weighted regression for nonparametric small area estimation, Technical Report 266, Dipartimento di Statistica e Matematica applicata all'Economia, Pisa.
- Särndal C.E., Swensson B. and Wretman J. (1992) *Model Assisted Survey Sampling*, Springer, Berlin, New York.
- Stehman S. and Overton W. (1989) Pairwise inclusion probability formulas in random-order, variable probability, systematic sampling, Technical Report 131, Department of Statistics, Oregon State University, Corvallis, OR.
- Stoddard J.L., Kahl J.S., Deviney F.A., DeWalle D.R., Driscoll C.T., Herlihy A.T., Kellog J.H., Murdoch P.S., Webb J.R. and Webster K.E. (2003) Response of surface water chemistry to the clean air act amendments of 1990, Technical Report EPA/620/R-93/001, U.S. Environmental Protection Agency.
- Wu C. (2003) Optimal calibration estimators in survey sampling, *Biometrika*, 90, 937–951.
- Wu C. (2005) Algorithms and R codes for the pseudo empirical likelihood method and the Rao-Sampford sampling method, *Survey Methodology*, to appear.
- Wu C. and Rao J.N.K. (2005) Pseudo empirical likelihood ratio confidence intervals for complex surveys, *Working paper*, 2004–06, Department of Statistics and Actuarial Science, University of Waterloo.
- Wu C. and Sitter R. (2001) A model-calibration to using complete auxiliary information from survey data, *Journal of the American Statistical Association*, 96, 185–193.